Metabolic engineering strategies in diatoms reveal unique phenotypes and genetic configurations with implications for algal genetics and synthetic biology

Jestin George[1], Tim Kahlke[1], Raffaela M. Abbriano[1], Unnikrishnan Kuzhiumparambil[1], Peter J. Ralph[1], Michele Fabris[1,2*]

[1] University of Technology Sydney, Climate Change Cluster, Broadway Campus, Ultimo NSW 2007, Australia

[2] CSIRO Synthetic Biology Future Science Platform, Brisbane, Qld 4001, Australia

*** Correspondence:**
Corresponding Author
michele.fabris@uts.edu.au

Format: British English

Keywords: Microalgae, *Phaeodactylum tricornutum*, extrachromosomal expression, random integration, long-read sequencing, integration islands, heterologous monoterpenoids, synthetic biology

Abstract

Diatoms are photosynthetic microeukaryotes that dominate phytoplankton populations and have increasing applicability in biotechnology. Uncovering their complex biology and elevating strains to commercial standards depends heavily on robust genetic engineering tools. However, engineering microalgal genomes predominantly relies on random integration of transgenes into nuclear DNA, often resulting in detrimental "position-effects" such as transgene silencing, integration into transcriptionally-inactive regions, and endogenous sequence disruption. With the recent development of extrachromosomal transgene expression via independent episomes, it is timely to investigate both strategies at the phenotypic and genomic level. Here, we engineered the model diatom *Phaeodactylum tricornutum* to produce the high-value heterologous monoterpenoid geraniol, which, besides applications as fragrance and insect repellent, is a key intermediate of high-value pharmaceuticals. Using high-throughput phenotyping we confirmed the suitability of episomes for synthetic biology applications and identified superior geraniol-yielding strains following random integration. We used third generation long-read sequencing technology to generate a complete analysis of all transgene integration events, including their genomic locations, and arrangements associated with high-performing strains at a genome-wide scale with subchromosomal detail, never before reported in any microalga. This revealed very large, highly concatenated insertion islands, offering profound implications on diatom functional genetics and next generation genome editing technologies, and is key for developing more precise genome engineering approaches in diatoms, including possible genomic safe harbour locations to support high transgene expression for targeted integration approaches. Furthermore, we have demonstrated that exogenous DNA is not integrated inadvertently into the nuclear genome of extrachromosomal-expression clones, an important characterisation of this novel engineering approach that paves the road to synthetic biology applications.

Deleted: ,
Deleted: ,
Deleted: transgene
Deleted: are

1 Introduction

Diatoms are a diverse group of unicellular Stramenopile microalgae that have received substantial attention for their ecological importance (Armbrust, 2009) and biotechnological potential (Huang & Daboussi, 2017). Newly developed genetic resources hold much promise for diatom functional genetics studies and have propelled the model pennate diatom *Phaeodactylum tricornutum* into the field of synthetic biology. Firstly, next-generation genetic engineering tools have now been established in *P. tricornutum*. These include targeted genetic engineering via transcription activator-like effector nucleases (TALENs) (Daboussi et al., 2014; Serif et al., 2017; Weyman et al., 2015) and CRISPR-Cas9 (Nymark et al., 2016; Serif et al., 2018; Sharma et al., 2018) techniques with wide applications for biotechnology and basic research. Next-generation engineering strategies also include the recently demonstrated extrachromosomal approach. Here, potentially large episomes that contain various DNA parts can be maintained and expressed without requiring genomic integration (Karas et al., 2015). Consequently, extrachromosomal transformation is anticipated to become increasingly widely-used in diatom genetic engineering (Huang & Daboussi, 2017). These next-generation engineering strategies are central to diatom genetics and synthetic biology primarily because they allow multi-gene stacking approaches (Ainley et al., 2013; Goyal et al., 2009). Secondly, the recently developed Universal Loop (uLoop) assembly kit provides a collection of useful parts for modular DNA assembly and high-throughput testing as well as more complex gene-stacking designs in *P. tricornutum* and other diatoms (Pollak et al., 2019). Altogether, these resources offer unparalleled potential of diatoms such as *P. tricornutum* compared to other model algal chassis.

Deleted:

Deleted:

Deleted: chasses

Further to these developments, intrinsic desirable biological traits have elevated this microbe as a promising alternative to well-established chasses, *Escherichia coli* and *Saccharomyces cerevisiae*. Such traits include its robustness and scalability for industrial-scale growth (Hamilton et al., 2015); and –unlike bacteria and yeast species– its ability to fix carbon via photosynthesis for cheaper culture conditions. There is also an increased availability of transcriptomic, metabolomic, and proteomic datasets for uncovering previously unknown traits in diatoms (Ashworth et al., 2016; Longworth et al., 2016; Remmers et al., 2018; Smith et al., 2019) including unique aspects with potential biotechnological relevance (Allen et al., 2011; Fabris et al., 2014; Fabris et al., 2012; Kroth et al., 2008).

P. tricornutum is poised to become a widely-used, reliable chassis organism and has been validated in various high interest biotechnological applications, including in the production of bioplastic precursor compounds (Hempel et al., 2011a), therapeutic antibodies (Hempel et al., 2011b; Hempel & Maier, 2012), biofuels (Yao et al., 2014) and nutritional supplements (Hamilton et al., 2014). However, these approaches have all relied on the first generation genetic engineering strategy of randomly integrated chromosomal expression (RICE) of exogenous DNA. While RICE has been crucial for generating a myriad of transgenic *P. tricornutum* strains, both for basic (Lavaud et al., 2012; Liu et al., 2016) and applied research (Hempel & Maier, 2012); it is beset with silencing issues resulting in low transgene expression and stability (Cerutti et al., 2011). This is largely attributed to the 'position effect' phenomenon, whereby transgenes integrate into regions in the genome that are unfavourable for transgene expression (Doron et al., 2016; Gangl et al., 2015; Huang & Daboussi, 2017), such as transcriptionally repressed regions (Elgin, 1996). In microalgal research, there is virtually no information regarding the mechanisms driving and regulating RICE. Uncovering this knowledge is important for better understanding of a strategy that is still widely used today, including for CRISPR-Cas9 targeted genome editing (Greiner et al., 2017; Hopes et al., 2016; Nymark et al., 2016).

In order for next-generation engineering tools to deliver a variety of synthetic biology applications in *P. tricornutum* and to replace RICE, they need to be better characterised. For example, even though targeted integration has been demonstrated in this species (Weyman et al., 2015), there are no known safe harbour loci–regions in the nuclear genome that facilitate stable, high transgene expression–identified in *P. tricornutum* or any other eukaryotic photosynthetic microbe. Similarly, non-integrative episomes are extremely appealing for synthetic biology applications as backbones of self-maintaining mini-chromosomes. However, there is little knowledge available regarding mechanisms of episomal maintenance (Diner et al., 2016) and expression, including the level of transgene expression that can be achieved by extrachromosomal expression (EE), and whether fragments of episomal DNA are inadvertently integrated into the nuclear genome. This is because EE technology has only recently been described in diatoms with limited examples of its use to express transgenes. Resolving these knowledge gaps will enable understanding of how different genetic engineering strategies may alter the diatom's biology and DNA integration patterns for developing more reliable next-generation engineering approaches required for complex synthetic biology.

Deleted: s

Given the developments for diatom synthetic biology, *P. tricornutum* is now being explored for its potential for heterologous terpenoid production (Vavitsas et al., 2018). *P. tricornutum* is a promising alternative chassis compared to the more widely used bacteria and yeast species, which require extensive engineering to increase relatively low flux to isoprenoid biosynthesis (Paddon et al., 2013; Wang et al., 2017; Bian et al., 2017; Zurbriggen et al., 2012). Until recently, *P. tricornutum* had only been metabolically engineered to produce heterologous terpenoids betulin and lupeol (D'Adamo et al., 2018) by RICE. However, we have since demonstrated the first use of EE for metabolic engineering in *P. tricornutum* (Fabris et al., 2020), by expressing a geraniol synthase from the medicinal plant *Catharanthus roseus* (EC 3.1.7.11, *CrGES*) for the heterologous production of geraniol. Geraniol is a commercially relevant monoterpenoid with a variety of applications as flavourant, fragrances, and insect repellent (Chen & Viljoen, 2010). Geraniol is also the first intermediate in the monoterpenoid indole alkaloids (MIAs) biosynthesis pathway, which in *C. roseus* leads to the synthesis of the very high-value products with pharmaceutical applications (Caputi et al., 2018; Van Moerkercke et al., 2013).

Deleted: (Fabris et al., unpublished, submitted)

Herein, with the aim of laying the basis for more sophisticated synthetic biology and metabolic engineering strategies, we generated and thoroughly profiled libraries of transgenic *P. tricornutum* cell lines engineered to produce geraniol via either EE or RICE. By adopting high-throughput phenotyping, we unveiled important intrinsic differences both between and within EE and RICE cell lines. This revealed a small selection of high-performing RICE strains, as well as a highly consistent transgene expression phenotype across EE strains. We used long-read DNA sequencing to interrogate the genomes of selected EE and RICE lines. Our results provide a complete analysis of all integration events, genomic locations, and transgene arrangements associated with high-performing RICE strains at both genome-wide and subchromosomal scale, never reported before in any microalga, and confirmed the non-integrative nature of EE.

Our findings are key for understanding the underexplored dynamics of EE in diatoms, to be used as the basis for gene-stacking based synthetic biology applications such as metabolic pathways and complex genetic circuit assembly and expression. They also highlight the importance of moving to next-generation genetic engineering strategies in *P. tricornutum* and lay the groundwork for identifying putative genomic safe harbour locations that support high transgene expression for targeted integration strategies.

Methods

Microbial strains and growth conditions

Phaeodactylum tricornutum CCAP1055/1 was grown in liquid ESAW (Berges et al., 2001) supplemented with 50 µg/mL zeocin (Invivogen, San Diego, CA, USA) where appropriate, under 100 µE m^{-2} s^{-1} light in 21 °C shaking at 95 rpm. *P. tricornutum* induction media was prepared following ESAW protocol but without any addition of phosphate. *Escherichia coli* was grown in Luria broth supplemented with 100 µg/mL ampicillin.

Cloning and genetic construct assembly

Plasmids and episomes were constructed using Gibson assembly cloning kit (New England Biolabs, Hitchin, UK). Plasmids were propagated in *E. coli* strain Top10 and purified by Monarch Plasmid Miniprep Kit (New England Biolabs, Hitchin, UK). PCR amplification was performed using Q5 high fidelity polymerase (New England Biolabs, Hitchin, UK) and PCR screening was performed using GoTaq Flexi DNA polymerase (Promega, Wisconsin, United States) according to the manufacturer's instructions. Plasmid coding sequences were validated by Sanger sequencing (Macrogen Korea, Seoul, Korea). Episomes *pPtPBR11_AP1p_CrGES-mVenus* (mVenus NCBI Accesssion: AAZ65844.1) and a control *pPtPBR11_AP1p-mVenus* are described in Fabris et al. (2020). Expression plasmid for chromosomal integration, *pUC19_AP1pCrGES-mVenus,* was ligated by Gibson assembly into the pUC19 cloning vector linearised with BamHI. Both the *AP1p-CrGES-mVenus-FCBPt* expression cassette and the *FCBPp-ShBle-FCBPt* zeocin resistant cassette were amplified from *pPtPBR11_AP1p_CrGES-mVenus* using 5'-tcgagctcggtacccgggCTAACAGGATTAGTGCAATTC-3' forward primer and 5'-aggtcgactAGACGAGCTAGTGTTATTC-3' reverse primer; and 5'-agctcgtctAGTCGACCTGCACATATG-3' forward primer and 5'-tgcaggtcgactctagagAGACGAGCTAGTGTTATTC-3' reverse primer respectively. Similarly, *pUC19_AP1p-mVenus* expression plasmid for genomic integration was ligated by Gibson assembly into the pUC19 cloning vector linearised with BamHI. The *AP1p-mVenus-FCBPt* expression cassette was amplified using the same primers for amplifying P1p-*CrGES-mVenus-FCBPt* expression cassette from *pPtPBR11_AP1-mVenus* episome. Sequences of the vectors used in this work are provided in Supplementary File 2.

Deleted: Fabris et al.(unpublished, submitted)

Diatom transformation and conjugation

P. tricornutum was transformed by biolistic bombardment using PDS-1000/He System with Hepta adapter (Bio-Rad) for nuclear genomic integration of plasmid DNA (Kroth, 2007). Afterward, the cell mixture was left to recover 12 hours before being scraped and plated onto fresh ½ ESAW 50 µg/mL zeocin agar plates and left for 4-5 weeks for single colonies to appear. *E. coli* containing *pTA-Mob* (Karas, Diner, et al., 2015) and *pPtPBR11_AP1p_CrGES-mVenus* or *pPtPBR11_AP1p-mVenus* plasmids (Fabris et al., 2020) were used for conjugation with *P. tricornutum* according to protocol described by Diner et al. (2016). The cell mixture was scraped and plated onto 3-5 fresh ½ ESAW zeocin agar plates and left for 10-15 days when single colonies appeared. Single colonies generated by biolistic bombardment and conjugation were picked and inoculated into individual wells of 96-well round bottom plates containing 200 µl of ESAW supplemented with 50 µg/mL zeocin. The EE and RICE generated cell lines were incubated at 21 °C with 100 µE m^{-2} s^{-1} light for 1 week to adjust to liquid growth, after which they were subcultured every 4 days. For high-throughput screening, cell lines were subcultured over three weeks in ESAW supplemented with 50 µg/mL zeocin (or without

Deleted: (Fabris et al., unpublished, submitted)

supplementation for stability analysis), and induced with phosphate-free ESAW for 24 hours before flow cytometry.

Flow cytometry and fluorescence-activated cell sorting (FACS)

Once antibiotic resistant single colonies were established in liquid culture, they were used to inoculate fresh 200 µL of ESAW with and without zeocin. Cell lines were subcultured 1:7 (v:v) every 4 days for 3 weeks under these conditions. On day 4 of culture, plates were centrifuged at 2300 g for 3 minutes to pellet cells. The supernatant was removed and cells were washed with 150 µL induction media twice before being resuspended in 200 µL of induction media and induced for 24 hrs. Induced cells were screened by flow cytometry using CytoFLEX S (Beckman Coulter). Fluorescence was excited using a 488 nm laser. mVenus fluorescence was detected using a 525/40 nm filter and chlorophyll fluorescence was detected using 690/50 nm filter. Compensation of chlorophyll channel was set to 0.3. The distribution of eight single colonies of wild type *P. tricornutum* cultured and induced in the same way as transgenic cell lines were used as controls to determine the mVenus auto fluorescence for screening. Only chorophyll positive cells were included in the analysis to account for cell debris and other background events.

The per cell mVenus fluorescence of 20,000 events was log normalised and violin plots were created in Python using the seaborn data visualization library (v. 0.9.0) (Allen et al., 2018). Cell lines selected for geraniol production analysis were sorted using BD Influx FACS (BD Biosciences). All cell lines were cultured and induced as described above prior to FACS. Yellow mVenus fluorescence was detected using a 488 nm laser for excitation and a 530/40 nm filter. Wild type *P. tricornutum* cultured and induced in the same way as the transformants, was used as a control to determine the mVenus auto fluorescence distribution. A preliminary screen of a pooled sample of the top eight RICE_mV transformants was used to define the mVenus positive gate which did not overlap with the wild type control. One thousand cells from each transformant cell line that fell into this gate were collected into 96-well round bottom plate wells containing 200 µL ESAW media. After sorting, cells were incubated for 1-2 weeks in 200 µL ESAW supplemented with zeocin 50 µg/mL in 96-well round bottom plate, after which they were subcultured as described previously.

Geraniol capture and analysis

Cell lines analysed for geraniol production were scaled up to 50 mL preculture in ESAW supplemented with zeocin 50 µg/mL. Precultures were used to inoculate 50 mL fresh ESAW media without zeocin in 250 mL shake flasks at 10,000 cells/mL density on Day 0. Cultures in late-exponential phase were induced in the presence of isopropyl myristate ($C_{17}H_{34}O_2$) to capture volatile monoterpenoids (Jiang et al., 2017). On Day 4 the total culture was collected and centrifuged at 3000 g for 4 minutes to pellet cells. Cells were resuspended in 4 mL induction media and washed twice in induction media before being resuspended in 30 mL induction media in fresh 250 mL shake flasks with 1.6 mL isopropyl myristate, which was harvested 72 hours after induction and stored -80 °C until being analysed by GCMS. Geraniol was captured, sampled and analysed as described in (Fabris et al., 2020).

Deleted: Fabris et al. (unpublished, submitted)

High molecular weight genomic DNA (gDNA) extraction

Genomic DNA from *P. tricornutum* transformants was extracted using $7x10^7$ cells in 4-6 extractions to obtain approximately 7.5 µg high molecular weight, purified gDNA (dx.doi.org/10.17504/protocols.io.qzudx6w). The DNA was resuspended in 25-45 µL ultrapure water overnight at room temperature.

Transgenic diatom phenotypes and genomes

MinION sequencing

MinION sequencing libraries were prepared according to the 1D Genomic DNA by ligation (SQK-LSK108) protocol supplied by the MinION manufacturer (Oxford Nanopore Technologies) with modifications. Briefly, the DNA fragmentation step was replaced with two bead-cleaning steps. An initial 1:0.1 bead clean of gDNA to GC Biotech CleanNGS (CNGS-0005) beads was performed and the sample was gently mixed by flicking, and slowly repeatedly inverted for 5 minutes, and pelleted on a magnetic rack to collect the supernatant. A second 1:1 bead clean was performed using the supernatant and beads. The sample was gently mixed by flicking, incubated by slowly rotating by hand for 5 minutes, and pelleted on a magnetic rack. The supernatant was removed and DNA bound to the beads was washed with 200 µL freshly prepared 70% ethanol twice without removing the tube off the magnet or disturbing the pellet. The beads and DNA were resuspended in 46 µL ultrapure water, incubated at room temperature for 5 minutes, the beads were pelleted on the magnet and the supernatant containing the DNA was collected and used according to the manufacturer protocol. Samples were sequenced until a coverage of seven to ten times was achieved. Raw reads were base called and quality filtered filtered using albacore v2.2.6 with default settings. All the sequencing data has been deposited in NCBI BioProject under the ID PRJNA593624.

Identifying transgene integration locations in nuclear genome

To identify reads which contain RICE plasmid DNA and remove reads made up of genomic DNA only, all reads were aligned against *pUC19_AP1p_CrGES-mVenus* RICE plasmid using BLAST. Reads that did align to RICE plasmid were defined as initial hits. To account for regions in *pUC19_AP1p_CrGES-mVenus* RICE plasmid that contain native *P. tricornutum* genomic regions, such as promoters or terminators, the *pUC19_AP1p_CrGES-mVenus* RICE plasmid sequence was aligned against *P. tricornutum* genome using BLAST (EnsemblProtists, ASM15095v2). Initial hits that only aligned to those regions of *pUC19_AP1p_CrGES-mVenus* that are native to *P. tricornutum* were filtered out as false-positive hits using custom awk commands. The resulting true-positive hits were manually checked to identify the chromosomes of the integration sites. For detailed analyses all reads were mapped to each of the matching chromosomes using bwa v0.7.15 (Li et al., 2009). The resulting sam files very sorted, converted to bam-format and indexed using samtools 1.3.1 (Li et al., 2009) and potential integration sites were manually checked using the Integrative Genomics Viewer v2.4.16 (Robinson et al., 2011). Further analysis for ambiguous hits was performed using the Artemis Comparison Tool (ACT) v13.0.0 (Carver et al., 2005).

Results and discussion

Terpenoid engineering in *P. tricornutum* has only recently been reported (D'Adamo et al., 2018; Fabris et al., 2020) and consequently, there is limited prior knowledge to inform metabolic engineering strategies in diatoms to obtain elevated terpenoid production. Recently, we demonstrated that extrachromosomal expression (EE) can be used to efficiently express the fusion protein CrGES-mVenus in *P. tricornutum* cytosol to produce up to 0.309 mg/L (0.21 µg/10^7cells) geraniol following bacterial conjugation (Fabris et al., 2020). EE of transgenes is not subject to position effect (Karas et al., 2015) and could therefore provide highly reproducible, consistent, and controllable expression, which is a basic requisite for synthetic biology. In contrast, randomly integrated chromosomal expression (RICE) can result in genetically dissimilar transformants and consequently varied transgene expression among them. However, diatom phenotypes derived from EE and RICE have not been systematically parameterised. Because little is known regarding the mechanisms and effects following EE and RICE of transgenes, it is unclear how these different engineering strategies will compare

Deleted: (D'Adamo et al., 2018; Fabris et al., unpublished, submitted)

Deleted: (Fabris et al., unpublished, submitted)

regarding the expression of *CrGES-mVenus*, and consequently, heterologous geraniol production. Therefore, we analysed the phenotypes of EE and RICE of *Ap1_CrGES-mVenus* in *P. tricornutum* cell lines both at the expression level and in terms of geraniol yield.

Two identical DNA expression cassettes of *AP1p_CrGES-mVenus* and *AP1p-mVenus* as control (Fabris et al., 2020), were cloned either into an pPtPBR11 episome (Diner et al., 2016) or a pUC19 plasmid (Norrander et al., 1983), and delivered either by bacterial conjugation or DNA-coated particle bombardment, respectively, in order to create EE and RICE *P. tricornutum* transformant libraries. Both *AP1p_CrGES-mVenus* constructs contained the *CrGES* gene fused at the carboxyl-terminus to a mVenus yellow fluorescent protein (YFP) (Kremers et al., 2006) for rapidly screening the cell lines by flow cytometry. The *CrGES-mVenus* fusion gene was driven by the *P. tricornutum* native alkaline phosphatase (*AP1*, *Phatr3_J49678*) promoter (hereafter *AP1p*), which is induced in low phosphate conditions for controllable expression (Lin et al., 2017).

Deleted: (Fabris et al., unpublished, submitted)

Upon transformation of *P. tricornutum* with the episomes and plasmids described above, the resulting antibiotic resistant cell lines were used to create four transgenic diatom libraries. Cell lines transformed with *pPtPBR11_AP1p_CrGES-mVenus* and *pPtPBR11_AP1p-mVENUS* for EE were denoted as EE_GmV and EE_mV, respectively. Likewise, cell lines transformed with *pUC19_AP1p_CrGES-mVenus* and *pUC19_AP1p-mVenus* for RICE were denoted as RICE_GmV and RICE_mV, respectively.

EE transformants demonstrate consistent mVenus fluorescence, while RICE transformants demonstrate higher, but more variable signals

In order to compare transgene expression across all four transgenic libraries (EE_GmV, EE_mV, RICE_GmV and RICE_mV), we required a high-throughput screening strategy to quantify the relative heterologous protein production. We used flow cytometry to rapidly evaluate the CrGES-mVenus expression in an unprecedented number of transformants to identify unique features among and within cell lines generated by EE and RICE. This strategy enabled us to confirm the correct expression of the fusion protein, as well as quantify its relative abundance, offering a proxy for identifying high-expressing transformant variants from the low expressing or silenced variants (Delvigne et al., 2014; Sheff & Thorn, 2004). It is generally accepted that RICE transformants will exhibit different levels of heterologous protein production from each other (Jeon et al., 2017; Tanwar et al., 2018; Hallmann, 2007). Conversely, EE exconjugants theoretically offer more consistent levels of expression (Karas et al., 2015). However, this has never been shown over a large scale of transformants or through direct comparison. Therefore, it is not known to what extent the expression of heterologous protein can vary across EE exconjugants or RICE transformants.

Single colonies from each of the four libraries were screened according to mVenus fluorescence to evaluate differences in transgene expression. Our results demonstrate that, as predicted and previously reported on a smaller sample (Fabris et al., 2020), EE results in consistent transgene expression among exconjugants. The mean mVenus fluorescence fold change of all EE_GmV lines and all EE_mV lines was 58.50- and 57.73- fold, respectively, when compared to WT auto-fluorescence and were not significantly different from each other (p>0.9999) (Figure 1A). This suggests that construct size and complexity –at least of this degree– does not affect expression, and confirms that EE might not be affected by variable transgene silencing typically associated with position effect. For EE_GmV strains, the DNA construct contained the fusion gene geraniol synthase and mVenus (total construct size of 10,849 bp); whereas EE_mV strains were transformed with the mVenus containing DNA (total construct size 9,082 bp). Within each EE library, there was also little variation among EE_GmV lines

Deleted: (Fabris et al., unpublished, submitted)

(SEM = 3.54) and EE_mV lines (SEM = 2.95) (Figure 1A). These results indicate that EE transformants are highly similar. Interestingly, every EE transformant analysed, across both EE_GmV and EE_mV libraries, showed a mean mVenus fluorescence greater than WT auto-fluorescence (Figure 1B). This shows that EE is highly efficient and reliable in generating cell lines that express transgenes and does not require extensive screening, as is required for RICE engineering strategies.

Unlike EE, RICE is subject to position effects and is therefore expected to result in transformants with variable transgene expression. We confirmed this in both RICE_GmV and RICE_mV libraries, in which transformants showed highly variable mVenus fluorescence intensities (Figure 1A). For example, the mean mVenus fluorescence in RICE_GmV lines ranged between 0.04- to 719.20- fold change (SEM = 21.37) and RICE_mV lines ranged between 0.40- to 1695.00- fold change (SEM = 43.22). Furthermore, the RICE_GmV and RICE_mV libraries were significantly different from each other ($p<0.0001$), demonstrating mean mVenus fold changes of 137.80- and 368.60-fold, respectively. Together these results suggested that when transgenes are integrated randomly in to the genome, features of the transgene, such as size and complexity, may affect its expression. It is plausible that gene silencing plays a role, as mVenus is present as a large fusion protein in the RICE_GmV library, whereas it is a smaller, free fluorescent protein in the RICE_mV library.

To further evaluate the heterogeneity of mVenus fluorescence across RICE libraries, we arbitrarily binned transformants based on the vast spread of mean mVenus fluorescence profiles recorded. We generated five groups comprising of less than 10-, 10- to 250-, 250- to 800-, 800- to 1500-, and greater than 1500-fold change in mVenus fluorescence compared to wild type auto-fluorescence. In the RICE_GmV and RICE_mV libraries (Figure 1B), approximately 25% of transformants showed less than 10-fold mean mVenus than wild type auto-fluorescence (Figure 1B). Biolistic bombardment is expected to result in random fragmentation of plasmid DNA (Hopes et al., 2016). This could theoretically result in antibiotic resistant transformants that contain the selection cassette without the intact Ap1*CrGES-mVenus* transcriptional unit, possibly resulting in antibiotic resistant transformants with fluorescence profiles indistinguishable from WT auto-fluorescence. Additionally, it is plausible that transformants associated with such low mean mVenus signals might have integrated the *CrGES-mVenus* expression cassette at transcriptionally repressed genomic loci, or in arrangements that may have triggered gene silencing (Kim et al., 2015). About 37% of RICE transformants and 100% of EE exconjugants demonstrated mean mVenus fluorescence 10- to 250-fold greater than wild type auto-fluorescence. A further 28% of RICE transformants showed a 250- to 800-fold increase. Only 7% showed an 800- to 1500-fold increase, and 1% (corresponding to 2 cell lines both from RICE_mV library) reached a remarkable 1500-fold increase in fluorescence compared to wild type auto-fluorescence. Together, these results demonstrate that transformants generated by RICE require extensive screening at the protein expression level, as up to a quarter can show no to low expression. Interestingly, RICE transformants were able to demonstrate exceptionally higher maximum transgene expression compared to EE using the same transgene cassette design. This warrants further investigation into these transgenic genomes, as there may be aspects of chromosome-integration that could be useful, particularly with regard to multi-generation transgene stability (Kohli & Christou, 2008). Although this high RICE-related expression could be advantageous for simple transgenic constructs, our results show that it would not be suitable for testing larger, more complex assemblies especially without reporter genes. Instead, the high, virtually size-independent consistency of phenotypes associated with EE offer a more suitable platform for applications involving multi-gene constructs, with the advantage of not requiring large scale screening.

Clonal variegation is broadly distributed in EE but discretely defined in RICE

Deleted: 2

Deleted: as

After having determined marked differences in the expression profile between EE and RICE libraries, we exploited the resolution of high-throughput flow cytometry to investigate the population composition within each cell line of EE and RICE libraries. In doing so, we identified relevant variations in the distributions of mVenus fluorescence within individual cell lines, known as cell mosaicism or variegation (Kaufman et al., 2008). Most EE transformants showed a relatively homogenous distribution of mVenus abundance within each cell line, such as that of EE_GmV-28, -47, -7 and -19, and EE_mV-72, -14 and -36 (Figure 1C and D). However, some showed increasingly diverse mVenus distribution profiles within individual cell lines, such as EE_GmV-92, -80, -33, -97 and -67 and EE_mV-22, -40, -5, -1, -65 and -97. Intriguingly, these cell lines tended to show higher mean mVenus abundance (Supp. Figure 1). This suggests that EE transformants which exhibited higher mean mVenus signals were composed of cells that were highly dissimilar from each other in a more continuous, non-discrete manner. This observation raises important questions about the dynamics of EE in diatoms; namely, how are episomal copies maintained within each cell, and how dynamic is episomal copy number and segregation across individual cells at different stages of the life cycle? Such mechanisms have been uncovered in other eukaryotic, non-microalgal species. For example, maintenance of viral episomes in mammalian cells and plasmids in *S. cerevisiae* have been attributed to chromosome tethering and hitchhiking mechanisms (Sau et al., 2019; Ghosh et al., 2007; McBride, 2008; Liu et al., 2008). In yeast synthetic biology research, episomal DNA sequences have been characterised based on traits including transformation efficiency, copy number, transgene expression and plasmid stability of clonal populations (Bouton & Smith, 1986; Gu et al., 2019; Nakamura et al., 2018). For example, Nakamura et al. (2018) tested various episome-regulation sequences in *Pichia pastoris* transgenic strains expressing EGFP extrachromosomally. They reported that strains containing the autonomously replicating sequence (ARS) without a centromeric region (CEN) showed broad fluorescence profiles similar to those that we report here, whereby cells within a single clonal population show a wide spread of EGFP fluorescence. However, when combined with centromeric region (CEN2), they reported a more discrete distribution of high EGFP fluorescence profiles that were more consistent with our RICE transformants. This was likely due to a strong bias for the mother cell over the daughter cells during cell division (Gehlen et al., 2011). While the pPTBR11 plasmid used in this study contains CEN region (Diner et al., 2016), it is still not yet known how such features contribute to episome expression in diatoms (Karas et al., 2015). Such factors could influence this cell-to-cell phenotypic heterogeneity, but for unknown reasons, this seemed to become more prominent at higher mean mVenus fluorescence.

Overall, RICE transformants demonstrated homogeneous fluorescence distribution profiles within individual cell lines, such as RICE_GmV-44, -41 and -64 and RICE_mV-93, -54, -50 and -4 (Figure 1E and F). Other RICE transformants also demonstrated heterogeneous mean mVenus profiles, but as numerous discrete populations within single cell lines (Figure 1E and F). For example, RICE_GmV-125, -120 and -3 and RICE_mV-24, -92, -2 and -85 all were composed of two unique populations of mVenus fluorescence distribution (Figure 1E and F). Transformant RICE_GmV-89 even showed three populations (Figure 1E). These results demonstrate that individual cells within a clonal transformant RICE cell line, generally assumed to have identical phenotypes, can be highly heterogeneous with regard to transgene expression, but that this heterogeneity can be distributed into unique, discrete populations. This is a major difference with the highly heterogeneous EE cell lines, which were instead characterized by a wide distribution of heterogeneity within the population, although RICE_GmV-68 transformant also followed this distribution.

Deleted: 1C

RICE cell lines exhibit dramatically varied stability that does not correlate to *CrGES-mVenus* expression

Transgenic diatom phenotypes and genomes

Given the extremely high outliers, we investigated the stability of the RICE libraries and how this related to expression level. Random chromosomal integration can result in stable maintenance and expression of transgenes, even in absence of selective pressure. Transformants that looked indistinguishable from wild type auto-fluorescence in the selective treatment did not change when selective pressure was removed (RICE_GmV-48 and RICE_mV-65, Figure 1G and H). We also identified RICE transformants that did not retain their mVenus fluorescence in absence of selective pressure (Figure 1G and H). For example, RICE_GmV-3 and -127 and RICE_mV-45, -59, -85 (Figure 1G and H, respectively) demonstrated a complete reduction in mVenus fluorescence when cultured in the absence of zeocin that was indistinguishable from wild type auto-fluorescence. Without selective pressure, cells that have silenced their resistance transgene (and by proxy, the transgene of interest) can outcompete and take over the culture due to the disadvantages of reduced energy and resource investment associated with transgene expression.

Interestingly, transgene expression level did not correlate to transgene stability, as seen in RICE_GmV-127 and -99, which showed similar expression levels with selection, but only -127 lost mVenus fluorescence when selection was removed. Likewise, RICE_GmV-3 and -127 cell lines show dissimilar mVenus signals in presence of selection but lost signal completely in selection-free conditions (Figure 1G and H). This suggested that there may be transgene integration events or arrangements that facilitate stable transgene expression and highlight the importance of designing screening procedures based on stability not only on transgene expression. Potential mechanisms of action include progressive transcriptional silencing via DNA methylation or histone modification, including de novo DNA methylation triggered by transgene recognition (Kohli et al., 2010); and posttranscriptional silencing known as RNA interference (Meyer, 1995; Cerutti et al., 2011; León-Bañares et al., 2004; Doron et al., 2016). Epigenetic silencing of nuclear-integrated exogenous DNA have been attributed to defense mechanisms against viruses and transposable elements in plants (Rajeevkumar et al., 2015) and mammalian cells (Alhaji et al., 2019) alike. Silencing mechanisms, and indeed transgene regulation mechanisms yet to be identified, can influence daughter cells from the same original clonal population differently (Kaufman et al., 2008). In fact, it is not known how stable randomly integrated exogenous DNA fragments are once they have been integrated, or how these insertions are genetically maintained over time.

In other RICE transformants, such as RICE_GmV-68, -33 and -10 and RICE_mV-17 and -70, we detected a reduction in mVenus abundance in absence of selection. Here, a distinct population of cells within each transformant showed signals similar to those in presence of selection, as well as a secondary population of noticeably lower mVenus fluorescence (Figure 1G and H). Other lines showed only a very small reduction in expression, such as RICE_76, -117 and -64 and RICE_mV-44, -94 and -25. Finally, we were able to identify some RICE transformants that maintained mVenus signals both in presence and absence of selective pressure, namely transformants RICE_GmV-99, -84, -24 and -47 and RICE_mV-66, -93 and -74 (Figure 1G and H). Some of these transformants also demonstrated comparatively high mVenus fluorescence abundance, particularly RICE_GmV-41, -47 and RICE_mV-74 (Supp. Figure 1). This suggests that they might contain integration events or arrangements that bypass silencing mechanisms. These transformants could provide empirical evidence for putative safe-harbour loci, which have been previously verified in various other organisms including mammalian cell lines (Cheng et al., 2016; Salsman & Dellaire, 2017; Papapetrou & Schambach, 2016; Lee et al., 2015), rice (Cantos et al., 2014) and cyanobacteria (Bentley et al., 2014; Pinto et al., 2015).

Together, these results once again highlight that RICE is not suitable for more complex synthetic biology and that efforts to move towards next-generation genetic engineering strategies is crucial. High expressing RICE transformants can be unstable, as well as contain unknown genomic disturbances and

mutations due to damage to the genome itself, as demonstrated in rice and maize (Liu et al., 2018). However, a better understanding of high transgene expression in RICE transformants may reveal aspects of exogenous DNA integration that would be useful for targeted insertion strategies.

Long-read whole-genome sequencing reveals no chromosomal integration of episomal DNA, whereas biolistic bombardment caused exogenous DNA to integrate at unique chromosomal loci

To date, transgenic genomic research has been restricted by prohibitive costs of whole genome sequencing and limited techniques that only reveal certain aspects of random integration events (Jeon et al., 2017; Scaife & Smith, 2016). In diatoms, such strategies include Southern blotting, which showed 1 - 10 transgene copies of foreign DNA integrated into the genome (Falciatore et al., 1999); quantitative polymerase chain reaction (qPCR), which revealed that copy number was relatively consistent between transformants variants (average of three) (D'Adamo et al., 2018); and thermal asymmetric interlaced PCR (TAIL-PCR), which revealed integration loci of exogenous DNA (Johansson et al., 2019). Similarly, the recently developed method of EE has been shown to require no transgene integration via episome recovery experiments (Diner et al., 2016; Karas et al., 2015). However, it is not yet known if integration does occur alongside extrachromosomal maintenance of episomes. Consequently, there is no knowledge of RICE or EE transgenic microalgal genomes with regard to exogenous DNA integration arrangements, the frequency of integration events throughout the genome, or any precise genomic integration loci. Answering some of these knowledge gaps is required for advancing synthetic biology design, progressing next-generation engineering tools –such as possible safe harbour loci to target–, and providing better understanding of the transgenic genome architecture, particularly regions associated with high transgene expression.

Developments in long-read sequencing technologies, namely Oxford Nanopore and PacBio, have allowed more continuous genome assemblies which can be done in real-time in the lab (Jain et al., 2018). To date, these technologies are mostly used for metagenomics analyses (Pinder et al., 2019; Robertsen et al., 2016) or sequencing new, non-model species (Davis et al., 2016; Fournier et al., 2017). Herein, we applied Oxford Nanopore sequencing to interrogate bacteria-conjugated and biolistic-bombarded transgenic *P. tricornutum* cell lines to explore integrated transgene arrangements, integration locations, and associated genetic architecture, as has been recently done in *Arabidopsis thaliana* and mouse models (Jupe et al., 2019; Nicholls et al., 2019). Given the phenotypic consistency between EE lines, we analysed a single EE line, EE_GmV-97, and assessed all the reads for alignment to the *pPtPBR11_AP1p_CrGES-mVenus* episome DNA. For RICE lines, we analysed two lines that showed high stability and mVenus fluorescence, RICE_GmV-41 and -47. These cell lines showed very similar transgene expression profiles and stabilities. Therefore, we aimed to identify differences or similarities regarding their transgenes at the genome-wide scale. These RICE reads were assessed for alignment to the RICE plasmid *pUC19_AP1p_CrGES-mVenus*. All EE and RICE reads with hits to their respective exogenous DNA constructs were then aligned to the wild type *P. tricornutum* genome in order to identify genomic integration events. The results of the Oxford Nanopore sequencing analysis are summarised in Table 1. For each cell line, we sequenced over 250 million nucleotides, resulting in genome coverage of five to nine times, with a probability greater than 99% that the complete genome of each cell line was covered (Clarke & Carbon, 1976).

Our results strongly suggest that no traces of episome were integrated into EE_GmV-97 (Table 1 and 2). Although we identified 26 reads that aligned to the episome, these reads contained no regions which aligned to the *P. tricornutum* genome (Table 1; Figure 2A), strongly suggesting that these DNA fragments did not get integrated into the nuclear chromosomes. This is the first demonstration that no episomal exogenous DNA is inadvertently integrated into the nuclear genome following bacterial

Deleted: reverse transcription

Deleted: RT

conjugation. Such knowledge is important for identifying any genetic disturbances that may go undetected in exconjugants, progressing knowledge for a better understanding of episomal regulation mechanisms, and for synthetic biology applications with *P. tricornutum* more broadly.

In the RICE lines, we identified only two independent integration loci in both RICE_GmV-41 and RICE_GmV-47, respectively (Table 1 and 2; Figure 2B and C). These four sites were detected and confirmed in both biological replicates (Table 2; Suppl. Figure 2). The integration events associated with these four independent loci consisted of concatenations of various fragments of the *pUC19_AP1p_CrGES-mVenus* RICE plasmid (Figure 2B and C). The genomic features and details associated with each of the four integration events are summarised in Table 2.

DNA extracted from each cell line does not come from a single cell, but instead a clonal population, and it is plausible that endogenous genomic regions around the insertion site are not stable (Kohli et al., 2006; Kohli et al., 1998). Therefore, it was not possible to resolve every integration event to the single nucleotide level, but only at < 60 bp range. For example, Kohli demonstrated that endogenous DNA concatenations can be assembled prior to or during integration in rice crop species, and suggested recombination could occur even after integration (Kohli et al., 2006; Kohli et al., 1998). It is also possible that insertions, deletions, or a combination of both (INDELs) can occur at the borders of an exogenous DNA integration event, driven by non-homologous integration (Shin et al., 2016). Such INDELs would cause reads at this small border region to show no alignment to wild type genome, as seen in RICE_GmV-47 integration event 47-10 (Suppl. Figure 2). Finally, the high sequencing error rate associated with Nanopore sequencing (15%) can also influence integration site determination.

In RICE_GmV-41, fragments of the *pUC19_AP1p_CrGES-mVenus* RICE plasmid were inserted at two unique genomic loci, ch1: 2,477,260 (integration island 41-1) and ch11: 316,959 – 317,016 (integration island 41-11) (Table 2). Both integration island 41-1 and 41-11 occur at intergenic regions in the genome; however, they are both flanked by predicted protein coding genes (Table 2; Figure 2B). Integration island 41-1 is situated 199 bp downstream of the 3' end of *Phatr3_J8770* and 479 bp upstream of the 5' start of *Phatr3_J54066* (Table 2; Figure 2B). *Phatr3_J8770* contains dynamin domains and *Phatr3_J54066* is putatively involved in vesicle trafficking functions according to HMMER (Finn et al., 2011) searches. Integration island 41-11 occurs ~900 bp downstream of the 3' end of *Phatr3_J46733* and ~100 bp upstream of the 5' start of *Phatr3_EG00809* (Table 2; Figure 2B). *Phatr3_J46733* contains a transmembrane feature at its C terminus (Uniprot) and a VAD1 Analog of StAR-related lipid transfer domain (VASt) according to HMMER (Finn et al., 2011). *Phatr3_EG00809* showed no predicted functional annotations, nor similarity to known protein domains (Finn et al., 2011).

Neither of these islands disrupted the protein coding regions of these neighbouring genes and we did not detect any growth defective phenotypes for these cell lines. However, the close proximity of the islands to these neighbouring genes means that the integration events may have affected their associated endogenous regulatory regions (Table 2, Figure 2B).

In transformant RICE_GmV-47, two integration events were localised to ch9: 865,083 - 865,119 (integration island 47-9) and ch10: 609,260 - 609,276 (integration island 47-10) (Table 2, Figure 2C). Both of these loci harbour predicted single-exon protein coding regions *Phatr3_J46300* and *Phatr3_J46528*, respectively, with no predicted functional annotations, nor similarity to known protein domains (Finn et al., 2011).

Interestingly, all four integration events were contained within unique sites across the entire genome of both cell lines, instead of occurring in a more scattered arrangement at a high number of locations, as has been demonstrated following biolistic bombardment in the plants *Oryza sativa* and *Zea mays* (Liu et al., 2018).

Biolistic bombardment results in extremely large integration islands containing highly repetitive arrangements of exogenous DNA

Due to the size of the *pUC19_AP1p_CrGES-mVenus* RICE plasmid (6.5 Kbp) and the length of the longer reads we obtained (up to 193.5 Kbp in length), we expected that single reads sequenced using this technology could span the entire integration site. We found this to be true for integration island 41-11, as demonstrated with right-left border read (LRB-R), 28.6 Kbp in length (Figure 2B). This read revealed ~12 Kbp aligned to the RICE plasmid and 4 Kbp upstream and 14 Kbp downstream of the 'integration island' aligning to adjacent loci in the reference genome (Figure 3A). This is the first visualisation of a complete integration event in any microalgae and clearly shows both the integration location in the genome as well as the integration event arrangement.

However, this was the only integration event which was detected on a single read. We did not expect that every other integration event would be so large that it would not be detectable within a single long read, but instead only align to one border of the integration island. This is demonstrated by representative reads 47-9_LB-R and 47-9_RB-R that contained a short 5' flank aligning from ch9: 860,407 - 865,083 and a 3' flank aligning to ch9: 865,119 - 867,673, respectively (Figure 3C). The majority portion of these two reads (42 Kbp for 47-9_LB-L and 82 Kbp for 47-9_RB-R) in fact aligned to RICE plasmid and were found to contain a high frequency of adjacent concatenations of *pUC19_AP1p_CrGES-mVenus* RICE plasmid, in both sense and antisense orientations (Figure 3C). This type of 'bordered' configuration was detected around the integration islands 41-1, 47-9, and 47-10.

The highly repetitive, shuffled structure of the integration islands may be responsible for the high *CrGES-mVenus* expression associated with these cell lines compared to others in this library. Alternatively or in addition to this, it is plausible that either high transcriptional intact unit copy number, or highly repetitive, chimeric promoter and/or terminator arrangements of the *CrGES-mVenus* cassette might act as enhancers for the expression of this construct. Given that we were able to detect mVenus fluorescence and geraniol production (Figure 4B and C), we can infer that some of these fragments contained functional transgene cassettes. However, the ~15% error rate of Nanopore sequencing technology (Jain et al., 2018) and the inability to detect single nucleotide polymorphisms make it near impossible to infer the exact number of functional copies present. Furthermore, it is unclear how stable these large integration islands are, if they are prone to recombination events, or what molecular mechanisms occurred to generate them.

Together, the left and right border reads for integration islands 41-1, 47-9 and 47-9 indicate that these islands are a minimum of 43 Kbp, 124 Kbp, and 87 Kbp, respectively. We then looked to assess whether these left and right border reads for these three integration islands aligned to each other to 'close' the integration island, but were unable to due to their repetitive nature (Figure 3D).

Interestingly, over 80% of the reads we identified that aligned to the *pUC19_AP1p_CrGES-mVenus* RICE plasmid did so in their entirety (Table 1); i.e. these reads contained no flanks which aligned to the genome at all (Figure 3B). Theoretically, these large reads with no genome-aligning flanks could sit between the left and right border reads of the integration islands. Although we cannot align these

highly concatenated reads to each other to confirm this, we speculate that islands 41-1, 47-9 and 47-10 could certainly be hundreds of kilobase pairs in size. Given that RICE_GmV-41 transformant has only two integration events and we knew the size of integration island 41-11 (~10 Kbp) and the sizes of all the 'RICE plasmid only' aligning reads together (1,808,244 nt), we can roughly estimate the size of integration island 41-1 through

$$\frac{\sum ni - (l_{i1} * c)}{c}$$

where *ni* is the number of nucleotides in 'RICE plasmid only' aligned reads, l_{il} is the length of integration island 41-11, and *c* is the estimated coverage. Following this the estimated length of the second integration island 41-1 is ~250 Kpbs. This correlates to 38 hypothetical back-to-back integrations of the full *pUC19_AP1p_CrGES-mVenus* RICE plasmid. Such large hypothetical islands are supported by similar results in transgenic rice and maize lines obtained following biolistic bombardment, in which large integration islands up to 1.6 Mbp in size were reported (Liu et al., 2018). Jupe et al. (2019) also used whole genome sequencing to elucidate transgene integration structure following *Agrobacterium*-mediated transformation in *Arabidopsis thaliana*. They reported between one and seven integration islands of between 20 and 230 Kbp per strain. While these results are from higher plant species, they confirm that these huge, highly concatenated islands are not specific to biolistic transformation, nor to diatoms.

As previously described, mechanisms involved with RICE are not well understood but have been investigated in plants (Kohli et al., 2006). Our results highlight a need to explore mechanisms driving the assembly and maintenance of these islands in diatoms, especially given the range of sizes possible that suggest numerous strategies may be at play. This is the first insight into how nuclear integration occurs in diatoms with widespread implications for existing understanding and future studies. Previous short read sequencing techniques such as targeted gene-walking (Parker et al., 1991), and inverse PCR or TAIL-PCR (Johansson et al., 2019; Liu & Chen, 2007; Huang et al., 2000) have been useful for identifying integration loci, but would not have been able to detect such large integration events of hundreds of kilobase pairs in size, nor would it be able to detect the complex transgene rearrangements (Nicholls et al., 2019) that we unveiled through long-read sequencing. Furthermore, highly concatenated integrations contain many repeated sequences (Figure 3), which nested primers used in TAIL-PCR are able to anneal to. Consequently, PCR strategies would result in many non-specific amplicons and difficulty in determining an unknown integration location. Whilst Southern blotting is a useful approach for determining gene copy number, it does not provide information about the integration site. Hence, our results demonstrate that long-read whole genome sequencing is an ideal, rapid and affordable approach for determining highly complex transgene integration events.

RICE *CrGES-mV* transformants are associated with higher expression and higher geraniol yield

We previously demonstrated that wild type *P. tricornutum* does not naturally produce geraniol, but that it can be efficiently engineered to extrachromosomally express *CrGES-mV* to produce it heterologously (Fabris et al., 2020). In order to determine whether extrachromosomal or chromosomal-integrated expression of the fusion construct *CrGES-mV* affect the heterologous production of monoterpenoids, we quantified the amount of geraniol produced in three independent RICE_GmV transformant lines and three EE_GmV exconjugant lines. Using mVenus fluorescence as a proxy for GES expression, we selected six transformants (RICE_GmV-24, -41, -24; and EE_GmV-97, -98 and -99, respectively) based on their mean fluorescence intensity and stability (Supp. Figure 1). With the aim of enriching the clonal populations associated with higher mean mVenus fluorescence, RICE_mV-74, RICE_GmV-

Deleted: 8

Deleted: and the configuration of the rearrangements as we report here following long-read sequencing.

Deleted: ¶

Deleted:

Deleted: (Fabris et al., unpublished, submitted)

24, -41 and -47 and EE_mV-97, EE_GmV-97, -98 and -99 were induced and sorted based on mVenus fluorescence using fluorescence activated cell sorting (FACS). A preliminary screen of a pooled sample of the top eight RICE_mV transformants was used to define the mVenus positive gate which did not overlap with the wild type control. During FACS, one thousand cells from each transformant that fell into this gate were collected and scaled up. We concluded that cell sorting did not enrich phenotypic populations, and in this specific case did not improve mean mVenus intensity (Supp. Figure 3).

Geraniol production in transgenic cell lines was evaluated in a bi-phasic, batch fermentation experiment (Fabris et al., 2020). All *CrGES-mV* transformants and exconjugants were induced by resuspension in lower volumes (30 ml) of phosphate-free media and showed similar growth to *mV* transformant and exconjugant controls prior to induction, indicating no identifiable loss of fitness in *CrGES-mV* expressing lines (Figure 4A).

We tracked mVenus fluorescence daily and quantified the accumulated geraniol 72 hours after inducing the expression of the *CrGES-mVenus* fusion gene. The RICE_GmV transformants, which showed increased mVenus fluorescence 24 hours after induction, produced more geraniol than the episomal exconjugant equivalents (Figure 4B and C). Chromosome-integrated transformant production yields were 150.9 ng/10^7 cells (0.37 mg/L), 304.4 ng/10^7 cells (0.89 mg/L) and 235.3 ng/10^7 cells (0.61 mg/L) for RICE_GmV-24, RICE_GmV-41 and RICE_GmV-47 respectively (Figure 4C). Conversely, non-integrated exconjugants demonstrated consistently lower yields that were 49.1 ng/10^7 cells (0.10 mg/L), 61.4 ng/10^7 cells (0.15 mg/L) and 72.5 ng/10^7 cells (0.15 mg/L) for EE_GmV-97, EE_GmV-98 and EE_GmV-99 respectively (Figure 4C). Neither EE nor RICE mVenus controls showed any detectable geraniol (Figure 4C). These results indicate that mVenus fluorescence is a reliable proxy for geraniol production, as mVenus fluorescence correlated with geraniol yields. Also, the high geraniol yields achieved support the hypothesis that *P. tricornutum* may have an available free pool of cytosolic geranyl diphosphate (GPP), the prenylphosphate precursor that CrGES converts into geraniol (Fabris et al., 2020), and that the heterologous synthesis of this monoterpenoid might not be limited by substrate availability in these settings. In light of these results, strategies involving promoter optimisation and targeted integration–both currently being evaluated in our laboratory–could further increase the geraniol production. Together, our results warrant the development of *P. tricornutum* for enhanced monoterpenoid production and suggest that it would be possible to improve production levels further by optimising *CrGES* expression at the genetic level.

Conclusions

Within the emerging application of monoterpenoid engineering in diatoms, we set out to generate specific knowledge to inform genetic optimisation strategies, including multi-gene approaches for synthetic biology and pathway engineering. We provided for the first time a comprehensive comparative analysis of two main types of transgene expression in diatoms, the conventional RICE and in the newly developed EE, using large-scale, high-throughput phenotyping, which allowed us to uncover details of these genetic resources between and within cell lines. The genetic differences between EE and RICE were reflected by the varied yields of the relevant monoterpenoid geraniol in a selection of transgenic diatom cell lines expressing a CrGES-mVenus fusion enzyme. The geraniol yield was more than four-fold higher in the best RICE transformant, reaching the titre of 304.4 ng/10^7 cells (0.89 mg/L), compared to the best EE exconjugants, reaching 72.47 ng/10^7 cells (0.15 mg/L). Thus, this work evaluated how previously unexplored genetic strategies can improve heterologous production of geraniol in *P. tricornutum*, in addition to more conventional strategies such as metabolic engineering or bioprocessing. While diatom engineering for terpenoid production has only recently

Deleted: (Fabris et al., unpublished, submitted)

Deleted: 1

Deleted: 3

Deleted: (Fabris et al., unpublished, submitted)

Deleted: (

been demonstrated, our results show that *P. tricornutum* is a promising photosynthetic microbial factory (D'Adamo et al., 2018; Fabris et al., 2020).

We report profound differences in the phenotypes of RICE and EE *P. tricornutum* cell lines in terms of expression levels, phenotypic consistency and sub-clonal population composition. Non-integrative episomes are associated with much more consistent phenotypes in the scenarios we tested regarding overexpression and EE exconjugants do not seem to require extensive screening. Furthermore, bacterial conjugation tends to result in more clones in a shorter amount of time than biolistic bombardment. Altogether, these results indicate that EE will be an invaluable resource for genetic parts validation and modular assembly, and even automation of the design-build-test-learn cycle. These aspects of more complex synthetic biology strategies are crucial for heterologous production of high-value products such as monoterpenoids. Our results highlighted the particularly limited knowledge available on key aspects of EE in diatoms. This included stability, copy number and segregation patterns, and episome re-arrangement, which we did not observe but has been reported (Slattery et al., 2018). Such characterisations still need to be addressed to fully exploit EE as a synthetic biology platform in diatoms.

In contrast, RICE cell lines are associated with high variability and overall higher expression levels, and by using large-scale screening it is possible to isolate particularly high expressing lines. These superior diatom cell lines bear highly concatenated arrangements of exogenous DNA, present as vast islands within or nearby predicted protein-coding genes. This raises a concern about this widespread method of generating transgenic diatom cell lines, as disrupting numerous protein coding regions can introduce unknown changes to *P. tricornutum* physiology that may not be easily detected. This is a particularly relevant issue in functional genetics studies involving overexpression, knock-down or knock-out constructs, which are traditionally delivered by biolistics, and randomly integrated in the genome of diatoms. On the contrary, it is not yet known if such large, highly concatenated integration events might be a factor in transgene stability and expression. In such scenario, RICE via biolistic bombardment, might be preferable over EE for obtaining high expressing cell lines. Finally, although it has been shown that high copy number and transgene tandem repeats can cause transcriptional silencing of transgene cassettes in other organisms (Kaufman et al., 2008; Moritz, Woltering, Becker, & Göpfert, 2016), our findings highlight the need to explore copy number and transgene arrangement optimisation in more detail, as this may well not be the case in *P. tricornutum*.

Whilst it is generally accepted that exogenous DNA delivered by biolistic bombardment randomly integrates in diatom chromosomes, the implications of this may have previously been overlooked, particularly at a time when CRISPR-Cas9 technology is being developed. While there is a general concern in CRISPR research to monitor and prevent off-target cutting by CRISPR-Cas9 itself, our results demonstrate that off-target effects from random integration of exogenous constructs such as vector backbone and DNA-encoded CRISPR-Cas9 components, could be just as much cause for concern. In this way, generating a precise knock-in or knock-out genotype by randomly integrating CRISPR-Cas9 components is suboptimal. As suggested by other works, (Sharma et al., 2018; Stukenberg et al., 2018) our findings clearly demonstrate the need to move towards non-integrative alternatives, such as episomal expression (Slattery et al., 2018) and ribonucleoprotein delivery (Serif et al., 2018).

This research also identified putative safe-harbour or neutral loci that could be tested for targeted integration in *P. tricornutum*. Neutral sites in cyanobacteria have been used for targeted integration in metabolic engineering for multigene pathway assembly (Bentley et al., 2014) and dual knock-in knock-out modifications (Li et al., 2016). Synthetic "landing pads" are useful for gene stacking via "domino

Deleted: (D'Adamo et al., 2018; Fabris et al., unpublished, submitted).

Deleted: However, these

Deleted: bare

Deleted: Furthermore,

Deleted: . However,

cloning", but depend on the knowledge of robust, reliable safe harbour loci prior to being feasibly applied to diatoms (Karas et al., 2015b). While some of the loci we identified harbour predicted protein coding regions, this is not unusual for safe harbours, as seen in human cell lines (e.g. CCR5 and ROSA26 loci) and mouse cell lines (e.g. Rosa26 locus), which all occur within protein coding regions. Furthermore, regions that may currently appear to be intergenic or non-functional may be re-categorised in the future, as more information about "junk DNA", transcripts without function (TUFs) and unannotated regulatory regions are discovered (Gingeras, 2007).

Finally, to the best of our knowledge, this work reports for the first time the suitability and utility of third generation long-read whole-genome sequencing to reveal the previously unknown nature of chromosomal integration sites, that would not have been feasible with conventional short-read sequencing. Future work investigating trans-genomes, such as low expression RICE cell lines or epigenetic modifications including DNA methylation patterns (Jain et al., 2018; Jupe et al., 2019), would build upon this knowledge to help uncover mechanisms driving transgene integration in diatoms. Such knowledge is important for developing better functional genomics tools including targeted genome editing. Our research primarily aimed at tracking specific, known transgenic constructs in EE and RICE transgenic diatoms cell lines. Long-read whole-genome sequencing technology can also be used to identify changes to the genome independent of an integration event, such as large translocations (Jupe et al., 2019) and deletions (Nicholls et al., 2019), purely due to the disruptive nature of the DNA delivery method. Our work lays the basis for future research efforts specifically focused on these relevant aspects, to investigate the impact of biolistic bombardment itself on genome integrity.

In conclusion, advancing synthetic pathway construction in *P. tricornutum* would ideally combine the reproducibility of EE with the high expression achievable through RICE, which could be achieved by targeted chromosomal integration. This work lays key groundwork for these developments which are crucial for extending knowledge on diatom biology and elevating model species such as *P. tricornutum* as a widely used synthetic biology chassis organism in a broad array of biotechnological applications.

Non-standard abbreviations

EE: Episomal expression

RICE: Randomly integrated chromosomal expression

Conflict of interest

None

Author contributions

JG designed the study, conducted the experiments and data analysis and drafted the manuscript; MF designed the study, conducted and supervised the experiments and drafted the manuscript; TK conducted the sequencing and data analysis and helped draft the manuscript; RA conducted lab experiments and helped draft the manuscript; PJR revised and edited the manuscript.

Funding

This work was funded by the University of Technology Sydney and the CSIRO Synthetic Biology Future Science Platform. JG is supported by a UTS Doctoral Scholarship. MF is supported by a CSIRO

777 Synthetic Biology Future Science Platform Fellowship co-funded by CSIRO and the University of
778 Technology Sydney.

779 **Acknowledgements**

780 The authors would like to thank Kun Xiao and Taya Lapshina for technical assistance.

781 **Data availability statement**

782 The datasets generated for this study can be found in NCBI BioProject repository under the ID
783 PRJNA593624.

784

Figure legends and tables

Figure 1. **mVenus fluorescence intensities of transgenic *P. tricornutum* extrachromosomal expression (EE) and randomly integrated chromosomal expression (RICE) libraries**. (A) Fold change of mean mVenus fluorescence of RICE_GmV and RICE_mV transformant libraries and EE_GmV and EE_mV libraries compared to wild type auto-fluorecence. Peach indicates CrGES-mVenus transgenic cell lines and teal indicates mVenus transgenic cell lines. Statistical comparisons were made using Kruskal-Wallis non-parametric ANOVA and Dunn's post-hoc test. For EE_GmV library n = 96 cell lines total, EE_mV library n = 96 cell lines total, RICE_GmV library n = 74 cell lines total and RICE_mV library n = 95 cell lines total. (B) Percentage of pooled RICE libraries (green) compared to percentage of pooled EE libraries (yellow) binned according to mean mVenus fluorescence fold change. (C-F) Violin plots indicate the per cell mVenus fluorescence intensity of ten representative cell lines for each library (C) EE_GmV (D) EE_mV (E) RICE_GmV (F) RICE_mV, ranked from lowest to mean mVenus expression (n = 20,000 cells for each cell line). (G and H) Representative cell lines from transgene stability analysis for (G) RICE_GmV and (H) RICE_mV libraries. Pink indicates selection free growth conditions, green indicates zeocin selection growth conditions, cell lines are ranked by mean mVenus intensity. (n = 20,000 cells for each cell line).

Table 1 Summarised details of MinION sequencing of EE and RICE transformants

	EE_ GmV-97	RICE_GmV-41_A	RICE_GmV-41_B	RICE_GmV-47_A	RICE_GmV-47_B
Nucleotides (total)	203,640,859	254,953,905	279,294,842	257,611,430	325,110,058
Reads (total)	26,455	37,373	31,581	18,525	20,153
Average read length	7,697.63	6,822	8,843.60	13,906	16,132.09
Coverage estimated	~5.8x	~ 7.2x	~7.9x	~ 7.3x	~9.2x
Total reads aligning to RICE plasmid	26	248	210	157	151
Reads aligning to RICE plasmid and genome on both borders	0	1	0	0	0
Reads aligning to RICE plasmid and genome on either border	0	18	14	19	27
Reads aligning to RICE plasmid only	26	230	196	138	124

Deleted: 0

808

809 **Table 2** Summarised details of integration events of EE and RICE transformants

Clone	Insertion site ID	Estimated chromosomal location (bp)	*In silico* assembly of island complete?	Size island (Kbp)	Genetic feature at integration site	Putative annotation of feature at integration site	AA size	Upstream features within 1kBp	Downstream features within 1kBp
EE_GmV-97	None				NA				
RICE_GmV-41	41-1	ch1: 2,477,260	Incomplete.	>43	Intergenic	None	NA	Phatr3_J8770 (protein coding)	Phatr3_J54066 (protein coding)
	41-11	ch11: 316,959 - 317,016	Complete. (Single read spanned island).	~10	Intergenic	None	NA	Phatr3_J46733 (protein coding)	Phatr3_EG00809 (protein coding)
RICE_GmV-47	47-9	ch9: 865,083 - 865,119	Incomplete.	>124	Phatr3_J46300 (single exon protein coding gene).	CM000612 Genomic DNA Translation: EEC47937.1	402 aa	NA	Phatr3_J46301 (protein coding)
	47-10	ch10: 609,260 – 609,276	Incomplete.	>87	Phatr3_J46528 (single exon protein coding gene).	CM000613 Genomic DNA Translation: EEC47687.1	429 aa	NA	NA

NA: Not applicable.

810

811 **Figure 2 Graphic representation of exogenous DNA constructs in extrachromosomal and**
812 **chromosomal DNA of the transgenic cell lines, based on long-read sequencing.** Only reads aligning
813 to both exogenous DNA and wild type *P. tricornutum* genome, and not those aligning to the genome
814 alone are depicted. (A) EE_GmV-97 transformant showed no reads which aligned to both exogenous
815 episomal DNA *pPtPBR11_AP1p-CrGES-mVenus* (green), and the reference *P. tricornutum* genome,
816 indicating that no exogenous DNA was integrated into the genome. Instead, some reads showed
817 alignment (red) only to episome DNA, suggesting these reads came from episomal DNA which was
818 extracted and analysed with genomic DNA. (B) Transformant RICE_GmV-41 generated by biolistic
819 bombardment showed reads which aligned to both exogenous RICE plasmid, *pUC19_AP1p-CrGES-*
820 *mVenus*, and the reference *P. tricornutum* genome, indicating a frequency of only two integration
821 islands, 41-1 and 41-11, occurring throughout the whole genome. Island 41-1 occurred on chromosome
822 1 where the longest left border read (LB-R) and right border read (RB-R) collectively indicated that
823 this island was a minimum of 43 Kbp in size. Island 41-11 occurred on chromosome 11 and was
824 spanned by a single read, right-left border read (RLB-R), which aligned to the reference genome at
825 both left and right borders (light blue), as well as the exogenous RICE plasmid. Red indicates alignment
826 in sense orientation and dark blue indicates alignment in antisense orientation, representing the highly
827 concatenated integration events observed. (C) RICE_GmV-47 transformant showed reads which
828 aligned to both exogenous RICE plasmid, *pUC19_AP1p-CrGES-mVenus*, and the reference *P.*
829 *tricornutum* genome, indicating a frequency of only two integration islands, 47-9 and 47-10, occurring

throughout the whole genome. Island 7-9 occurred on chromosome 9 where the longest left border read (LB-R) and right border read (RB-R) collectedly indicate that this island is a minimum of 117 kB in size. Island 47-10 occurred on chromosome 10 where the longest left border read (LB-R) and right border read (RB-R) collectedly indicate that this island is a minimum of 87 kB in size.

Figure 3 Graphic representation of rearrangements of exogenous DNA in *P. tricornutum* chromosomes, based on long-read sequencing. Red channels show alignment in sense orientation and blue channels show alignment in antisense orientation. Regions that are not highlighted did align to the plasmid, but with below-threshold for hit length of percent identity used for the visualisation, which was performed manually. (A) Alignment of a right-left border read (RLB-Read) (top) from integration event 41-11 to RICE plasmid *pUC19_AP1pCrGES-mVENUS* (bottom) and to the wild type *P. tricornutum* genome (green). (B) A single 97.5 Kbp read (bottom) with no regions of similarity to the *P. tricornutum* wild type reference genome aligned to the RICE plasmid *pUC19_AP1pCrGES-mVENUS* (top). (C) Integration island 47-9 made up by two reads; the left border read (LB-Read) (middle) contains approximately 42 Kbp aligned to the RICE plasmid (bottom) and 3 Kbp aligned to the *P. tricornutum* wild type reference genome (top). The right border read (RB-Read) (middle) contains approximately 82 Kbp of aligned to the RICE plasmid and 2 Kbp aligned to the *P. tricornutum* wild type reference genome. (D) Alignments of left and right border reads to each other for integration island 41-1, 47-9, and 47-10. These reads do not align to each other to 'close' the integration island, suggesting that some 'filler' reads are missing.

Figure 4 Geraniol production in three selected RICE_GmV and EE_GmV cell lines. N = 3, error bars represent SEM, statistical comparisons were made using one-way ANOVA and Tukey's multiple comparisons post-hoc test. (A) Growth curve for all cell lines. (B) mVenus fluorescence intensity 24 hours after induction. (C) geraniol produced after 72 hours induction.

References

Ainley, W. M., Sastry-Dent, L., Welter, M. E., Murray, M. G., Zeitler, B., Amora, R., … Petolino, J. F. (2013). Trait stacking via targeted genome editing. *Plant Biotechnology Journal*, *11*(9), 1126–1134. https://doi.org/10.1111/pbi.12107

Ajikumar, P. K., Xiao, W. H., Tyo, K. E. J., Wang, Y., Simeon, F., Leonard, E., … Stephanopoulos, G. (2010). Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli. *Science*, *330*(6000), 70–74. https://doi.org/10.1126/science.1191652

Alhaji, S. Y., Ngai, S. C., & Abdullah, S. (2019). Silencing of transgene expression in mammalian cells by DNA methylation and histone modifications in gene therapy perspective. *Biotechnology and Genetic Engineering Reviews*, *35*(1), 1–25. https://doi.org/10.1080/02648725.2018.1551594

Allen, A. E., Dupont, C. L., Oborník, M., Horák, A., Nunes-Nesi, A., McCrow, J. P., … Bowler, C. (2011). Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature*, *473*(7346), 203–207. https://doi.org/10.1038/nature10074

Allen, M., Poggiali, D., Whitaker, K., & Marshall, T. R. (2018). Raincloud plots: a multi-platform tool for robust data visualization. https://doi.org/10.7287/peerj.preprints.27137v1

Andrew E. Allen, Julie LaRoche, Uma Maheswari, Markus Lommer, Nicolas Schauer, Pascal J. Lopez, Giovanni Finazzi, Alisdair R. Fernie, and C. B. (2008). Whole cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation_2008_Proc Nat Acad Sci 105_10438-10443.pdf. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 10438–10443.

Armbrust, E. V. (2009). The life of diatoms in the world's oceans. *Nature*, *459*(7244), 185–192. https://doi.org/10.1038/nature08057

Ashworth, J., Turkarslan, S., Harris, M., Orellana, M. V., & Baliga, N. S. (2016). Pan-transcriptomic analysis identifies coordinated and orthologous functional modules in the diatoms Thalassiosira pseudonana and Phaeodactylum tricornutum. *Marine Genomics*, *26*, 21–28. https://doi.org/10.1016/j.margen.2015.10.011

Bentley, F. K., Zurbriggen, A., & Melis, A. (2014). Heterologous expression of the mevalonic acid pathway in cyanobacteria enhances endogenous carbon partitioning to isoprene. *Molecular Plant*, *7*(1), 71–86. https://doi.org/10.1093/mp/sst134

Berges, J. A., Franklin, D. J., & Harrison, P. J. (2001). Evolution of an artificial seawater medium: Improvements in enriched seawater, artificial water over the last two decades. *Journal of Phycology*, *37*(6), 1138–1145. https://doi.org/10.1046/j.1529-8817.2001.01052.x

Bian, G., Deng, Z., & Liu, T. (2017). Strategies for terpenoid overproduction and new terpenoid discovery. *Current Opinion in Biotechnology*, *48*, 234–241. https://doi.org/10.1016/j.copbio.2017.07.002

Bouton, A. H., & Smith, M. M. (1986). Fine-structure analysis of the DNA sequence requirements for autonomous replication of Saccharomyces cerevisiae plasmids. *Molecular and Cellular Biology*, *6*(7), 2354–2363. https://doi.org/10.1128/mcb.6.7.2354

Bozarth, A., Maier, U. G., & Zauner, S. (2009). Diatoms in biotechnology: Modern tools and applications. *Applied Microbiology and Biotechnology*, *82*(2), 195–201. https://doi.org/10.1007/s00253-008-1804-8

Cantos, C., Francisco, P., Trijatmiko, K. R., Slamet-Loedin, I., & Chadha-Mohanty, P. K. (2014). Identification of "safe harbor" loci in indica rice genome by harnessing the property of zinc-finger nucleases to induce DNA damage and repair. *Frontiers in Plant Science*, *5*(June), 302. https://doi.org/10.3389/fpls.2014.00302

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., & Parkhill, J. (2005). ACT: The Artemis comparison tool. *Bioinformatics*, *21*(16), 3422–3423. https://doi.org/10.1093/bioinformatics/bti553

Cerutti, H., Ma, X., Msanne, J., & Repas, T. (2011). RNA-mediated silencing in Algae: Biological roles and tools for analysis of gene function. *Eukaryotic Cell*, *10*(9), 1164–1172. https://doi.org/10.1128/EC.05106-11

Chen, W., & Viljoen, A. M. (2010). Geraniol - A review of a commercially important fragrance material. *South African Journal of Botany*, *76*(4), 643–651. https://doi.org/10.1016/j.sajb.2010.05.008

907 Cheng, J. K., Lewis, A. M., Kim, D. S., Dyess, T., & Alper, H. S. (2016). Identifying and retargeting
908 transcriptional hot spots in the human genome. *Biotechnology Journal*, 1100–1109.
909 https://doi.org/10.1002/biot.201600015.Submitted

910 Clarke, L., & Carbon, J. (1976). A colony bank containing synthetic Col El hybrid plasmids
911 representative of the entire E. coli genome. 1976. *Biotechnology (Reading, Mass.)*,
912 *9*(September), 91–99.

913 D'Adamo, S., Schiano di Visconte, G., Lowe, G., Szaub-Newton, J., Beacham, T., Landels, A., …
914 Matthijs, M. (2018). Engineering The Unicellular Alga Phaeodactylum tricornutum For High-
915 Value Plant Triterpenoid Production. *Plant Biotechnology Journal*, 0–2.
916 https://doi.org/10.1111/pbi.12948

917 Daboussi, F., Leduc, S., Maréchal, A., Dubois, G., Guyot, V., Perez-Michaut, C., Amato, A.,
918 Falciatore, A., Juillerat, A., Beurdeley, M., Voytas, D., Cavarec, L., & Duchateau, P. (2014).
919 Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology.
920 *Nature Communications*, *5*(May), 1–7. https://doi.org/10.1038/ncomms4831

921 Davis, A. M., Iovinella, M., James, S., Robshaw, T., Dodson, R., Herrero-davila, L., … Davis, S. J.
922 (2016). Using MinION nanopore sequencing to generate a. *Bioarxiv*.

923 Delvigne, F., Zune, Q., Lara, A. R., Al-Soud, W., & S??rensen, S. J. (2014). Metabolic variability in
924 bioprocessing: Implications of microbial phenotypic heterogeneity. *Trends in Biotechnology*,
925 *32*(12), 608–616. https://doi.org/10.1016/j.tibtech.2014.10.002

926 Diner, R. E., Bielinski, V. A., Dupont, C. L., Allen, A. E., & Weyman, P. D. (2016). Refinement of
927 the Diatom Episome Maintenance Sequence and Improvement of Conjugation-Based DNA
928 Delivery Methods. *Frontiers in Bioengineering and Biotechnology*, *4*(August).
929 https://doi.org/10.3389/fbioe.2016.00065

930 Doron, L., Segal, N., & Shapira, M. (2016). Transgene Expression in Microalgae — From Tools to
931 Applications Technical Approaches Used for, *7*(April), 1–24.
932 https://doi.org/10.3389/fpls.2016.00505

933 Elgin, S. C. R. (1996). Heterochromatin and gene regulation in Drosophila. *Current Opinion in*
934 *Genetics and Development*, *6*(2), 193–202. https://doi.org/10.1016/S0959-437X(96)80050-5

935 Fabris, M., George, J., Kuzhiumparambil, U., Lawson, C. A., Jaramillo Madrid, A. C., Abbriano, R.
936 M., Vickers, C., & Ralph, P. (2020). Extrachromosomal genetic engineering of the marine
937 diatom Phaeodactylum tricornutum enables the heterologous production of monoterpenoids.
938 *ACS Synthetic Biology*. https://doi.org/10.1021/acssynbio.9b00455

939 Fabris, M., Matthijs, M., Carbonelle, S., Moses, T., Pollier, J., Dasseville, R., Baart, J.E., Vyverman,
940 W. & Goossens, A. (2014). Tracking the sterol biosynthesis pathway of the diatom
941 Phaeodactylum tricornutum. *The New Phytologist*, 521–535. https://doi.org/10.1111/nph.12917

942 Fabris, M., Matthijs, M., Rombauts, S., Vyverman, W., Goossens, A., & Baart, G. J. E. (2012). The
943 metabolic blueprint of Phaeodactylum tricornutum reveals a eukaryotic Entner-Doudoroff
944 glycolytic pathway. *Plant Journal*, *70*(6), 1004–1014. https://doi.org/10.1111/j.1365-
945 313X.2012.04941.x

Deleted: …

Deleted: …

Deleted: …

949 Falciatore, A., Casotti, R., Leblanc, C., Abrescia, C., & Bowler, C. (1999). Transformation of
950 nonselectable reporter genes in marine diatoms. *Marine Biotechnology*, *1*(3), 239–251.
951 https://doi.org/10.1007/PL00011773

952 Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity
953 searching. *Nucleic Acids Research*, *39*(SUPPL. 2), 29–37. https://doi.org/10.1093/nar/gkr367

954 Fournier, T., Gounot, J. S., Freel, K., Cruaud, C., Lemainque, A., Aury, J. M., … Friedrich, A.
955 (2017). High-quality de novo genome assembly of the Dekkera bruxellensis yeast using
956 Nanopore MinION sequencing. *G3: Genes, Genomes, Genetics*, *7*(10), 3243–3250.
957 https://doi.org/10.1534/g3.117.300128

958 Gangl, D., Zedler, J. A. Z., Rajakumar, P. D., Martinez, E. M. R., Riseley, A., W??odarczyk, A., …
959 Robinson, C. (2015). Biotechnological exploitation of microalgae. *Journal of Experimental*
960 *Botany*, *66*(22), 6975–6990. https://doi.org/10.1093/jxb/erv426

961 Gehlen, L. R., Nagai, S., Shimada, K., Meister, P., Taddei, A., & Gasser, S. M. (2011). Nuclear
962 geometry and rapid mitosis ensure asymmetric episome segregation in yeast. *Current Biology*,
963 *21*(1), 25–33. https://doi.org/10.1016/j.cub.2010.12.016

964 Ghosh, S. K., Hajra, S., & Jayaram, M. (2007). Faithful segregation of the multicopy yeast plasmid
965 through cohesin-mediated recognition of sisters. *Proceedings of the National Academy of*
966 *Sciences of the United States of America*, *104*(32), 13034–13039.
967 https://doi.org/10.1073/pnas.0702996104

968 Gingeras, T. (2007). Origin of phenotypes: Genes and transcripts. *Genome Research*, (408), 682–
969 690. https://doi.org/10.1101/gr.6525007

970 Goyal, A., Bhowmik, P. K., & Basu, S. K. (2009). Minichromosomes : The second generation
971 genetic engineering tool. *Plant Omics Journal*, *2*(1), 1–8.

972 Greiner, A., Kelterborn, S., Evers, H., Kreimer, G., Sizova, I., & Hegemann, P. (2017). Targeting of
973 Photoreceptor Genes in Chlamydomonas reinhardtii via Zinc- finger Nucleases and
974 CRISPR/Cas9. *Plant Cell Advance Publication. Published on October*, *4*.
975 https://doi.org/10.1105/tpc.17.00659

976 Gu, Y., Gao, J., Cao, M., Dong, C., Lian, J., Huang, L., … Xu, Z. (2019). Construction of a series of
977 episomal plasmids and their application in the development of an efficient CRISPR/Cas9 system
978 in Pichia pastoris. *World Journal of Microbiology and Biotechnology*, *35*(6), 1–10.
979 https://doi.org/10.1007/s11274-019-2654-5

980 Hallmann, A., & Hallman, a. (2007). Algal transgenics and biotechnology. *Transgenic Plant J*, *1*(1),
981 81–98. https://doi.org/10.3390/ijms12010633

982 Hamilton, M. L., Haslam, R. P., Napier, J. A., & Sayanova, O. (2014). Metabolic engineering of
983 Phaeodactylum tricornutum for the enhanced accumulation of omega-3 long chain
984 polyunsaturated fatty acids. *Metabolic Engineering*, *22*, 3–9.
985 https://doi.org/10.1016/j.ymben.2013.12.003

986 Hamilton, M. L., Warwick, J., Terry, A., Allen, M. J., Napier, J. A., & Sayanova, O. (2015). Towards

the industrial production of omega-3 long chain polyunsaturated fatty acids from a genetically modified diatom phaeodactylum tricornutum. *PLoS ONE*, *10*(12), 1–15. https://doi.org/10.1371/journal.pone.0144054

Hempel, F., Bozarth, A. S., Lindenkamp, N., Klingl, A., Zauner, S., Linne, U., … Maier, U. G. (2011). Microalgae as bioreactors for bioplastic production. *Microbial Cell Factories*, *10*(1), 81. https://doi.org/10.1186/1475-2859-10-81

Hempel, F., Lau, J., Klingl, A., & Maier, U. G. (2011). Algae as protein factories: Expression of a human antibody and the respective antigen in the diatom phaeodactylum tricornutum. *PLoS ONE*, *6*(12). https://doi.org/10.1371/journal.pone.0028424

Hempel, F., & Maier, U. G. (2012). An engineered diatom acting like a plasma cell secreting human IgG antibodies with high efficiency. *Microbial Cell Factories*, *11*(1), 1. https://doi.org/10.1186/1475-2859-11-126

Hopes, A., Nekrasov, V., Kamoun, S., & Mock, T. (2016). Editing of the urease gene by CRISPR-Cas in the diatom Thalassiosira pseudonana. *Plant Methods*, 1–12. https://doi.org/10.1101/062026

Huang, G., Zhang, L., & Birch, R. G. (2000). Rapid amplification and cloning of Tn5 flanking fragments by inverse PCR. *Letters in Applied Microbiology*, *31*(2), 149–153. https://doi.org/10.1046/j.1365-2672.2000.00781.x

Huang, W., & Daboussi, F. (2017). Genetic and metabolic engineering in diatoms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1728), 20160411. https://doi.org/10.1098/rstb.2016.0411

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345. https://doi.org/10.1038/nbt.4060

Jeon, S., Lim, J.-M., Lee, H.-G., Shin, S.-E., Kang, N. K., Park, Y.-I., … Chang, Y. K. (2017). Current status and perspectives of genome editing technology for microalgae. *Biotechnology for Biofuels*, *10*(1), 267. https://doi.org/10.1186/s13068-017-0957-z

Jiang, G. Z., Yao, M. D., Wang, Y., Zhou, L., Song, T. Q., Liu, H., … Yuan, Y. J. (2017). Manipulation of GES and ERG20 for geraniol overproduction in Saccharomyces cerevisiae. *Metabolic Engineering*, *41*(March), 57–66. https://doi.org/10.1016/j.ymben.2017.03.005

Johansson, O. N., Töpel, M., Pinder, M. I. M., Kourtchenko, O., Blomberg, A., Godhe, A., & Clarke, A. K. (2019). Skeletonema marinoi as a new genetic model for marine chain-forming diatoms. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-019-41085-5

Jupe, F., Rivkin, A. C., Michael, T. P., Zander, M., Motley, S. T., Sandoval, J. P., … Ecker, J. R. (2019). The complex architecture and epigenomic impact of plant T-DNA insertions. *PLoS Genetics*, *15*(1), 1–25. https://doi.org/10.1371/journal.pgen.1007819

Karas, B. J., Diner, R. E., Lefebvre, S. C., McQuaid, J., Phillips, A. P. R., Noddings, C. M., … Weyman, P. D. (2015). Designer diatom episomes delivered by bacterial conjugation. *Nature*

Communications, *6*, 6925. https://doi.org/10.1038/ncomms7925

Karas, B. J., Suzuki, Y., & Weyman, P. D. (2015). Strategies for cloning and manipulating natural and synthetic chromosomes. *Chromosome Research*, *23*(1), 57–68. https://doi.org/10.1007/s10577-014-9455-3

Kaufman, W. L., Kocman, I., Agrawal, V., Rahn, H. P., Besser, D., & Gossen, M. (2008). Homogeneity and persistence of transgene expression by omitting antibiotic selection in cell line isolation. *Nucleic Acids Research*, *36*(17). https://doi.org/10.1093/nar/gkn508

Kim, E. J., Ma, X., & Cerutti, H. (2015). Gene silencing in microalgae: Mechanisms and biological roles. *Bioresource Technology*, *184*, 23–32. https://doi.org/10.1016/j.biortech.2014.10.119

Kohli, A., & Christou, P. (2008). Stable transgenes bear fruit. *Nature Biotechnology*, *26*(6), 653–654. https://doi.org/10.1038/nbt0608-653

Kohli, A., González-Melendi, P., Abranches, R., Capell, T., Stoger, E., & Christou, P. (2006a). The quest to understand the basis and mechanisms that control expression of introduced transgenes in crop plants. *Plant Signaling and Behavior*, *1*(4), 185–195. https://doi.org/10.4161/psb.1.4.3195

Kohli, A., González-Melendi, P., Abranches, R., Capell, T., Stoger, E., & Christou, P. (2006b). The quest to understand the basis and mechanisms that control expression of introduced transgenes in crop plants. *Plant Signaling and Behavior*, *1*(4), 185–195. https://doi.org/10.4161/psb.1.4.3195

Kohli, A., Leech, M., Vain, P., Laurie, D. A., & Christou, P. (1998). Transgene organization in rice engineered through direct DNA transfer supports a two-phase integration mechanism mediated by the establishment of integration hot spots. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(12), 7203–7208. https://doi.org/10.1073/pnas.95.12.7203

Kohli, A., Miro, B., & Twyman, R. M. (2010). *Transgene Integration , Expression and Stability in Plants : Strategies for Improvements*. https://doi.org/10.1007/978-3-642-04809-8

Kremers, G. J., Goedhart, J., Van Munster, E. B., & Gadella, T. W. J. (2006). Cyan and yellow super fluorescent proteins with improved brightness, protein folding, and FRET förster radius. *Biochemistry*, *45*(21), 6570–6580. https://doi.org/10.1021/bi0516273

Kroth, P. G. (2007). *Genetic transformation: a tool to study protein targeting in diatoms*. *Journal of Chemical Information and Modeling* (Vol. 53). Methods Molecular Biology 390. https://doi.org/10.1017/CBO9781107415324.004

Kroth, P. G., Chiovitti, A., Gruber, A., Martin-Jezeque, V., Mock, T., Parker, M. S., … Bowler, C. (2008). A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis. *PLoS ONE*, *3*(1). https://doi.org/10.1371/journal.pone.0001426

Lavaud, J., Materna, A. C., Sturm, S., Vugrinec, S., & Kroth, P. G. (2012). Silencing of the violaxanthin de-epoxidase gene in the diatom Phaeodactylum tricornutum reduces diatoxanthin synthesis and non-photochemical quenching. *PLoS ONE*, *7*(5).

https://doi.org/10.1371/journal.pone.0036806

Lee, J. S., Kallehauge, T. B., Pedersen, L. E., & Kildegaard, H. F. (2015). Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway. *Scientific Reports*, *5*, 1–11. https://doi.org/10.1038/srep08572

León-Bañares, R., González-Ballester, D., Galván, A., & Fernández, E. (2004). Transgenic microalgae as green cell-factories. *Trends in Biotechnology*, *22*(1), 45–52. https://doi.org/10.1016/j.tibtech.2003.11.003

Li, Heng, & Durbin, R. (2009). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, Heng, Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Hung, Shen, C. R., Huang, C.-H., Sung, L.-Y., Wu, M.-Y., & Hu, Y.-C. (2016). CRISPR-Cas9 for the Genome Engineering of Cyanobacteria and Succinate Production. *Metabolic Engineering*, *38*(August), 293–302. https://doi.org/10.1016/j.ymben.2016.09.006

Lin, H. Y., Yen, S. C., Kuo, P. C., Chung, C. Y., Yeh, K. L., Huang, C. H., … Lin, H. J. (2017). Alkaline phosphatase promoter as an efficient driving element for exogenic recombinant in the marine diatom Phaeodactylum tricornutum. *Algal Research*, *23*, 58–65. https://doi.org/10.1016/j.algal.2017.01.007

Liu. (2018). Genome-scale sequence disruption following biolistic transformation in rice and maize.

Liu, X., Hempel, F., Stork, S., Bolte, K., Moog, D., Heimerl, T., … Zauner, S. (2016). Addressing various compartments of the diatom model organism Phaeodactylum tricornutum via sub-cellular marker proteins. *Algal Research*, *20*, 249–257. https://doi.org/10.1016/j.algal.2016.10.018

Liu, Y. G., & Chen, Y. (2007). High-efficiency thermal asymmetric interlaced PCR for amplification of unknown flanking sequences. *BioTechniques*, *43*(5), 649–656. https://doi.org/10.2144/000112601

Longworth, J., Wu, D., Huete-Ortega, M., Wright, P. C., & Vaidyanathan, S. (2016). Proteome response of Phaeodactylum tricornutum, during lipid accumulation induced by nitrogen depletion. *Algal Research*, *18*, 213–224. https://doi.org/10.1016/j.algal.2016.06.015

Lorenzo Caputi, Jakob Franke, Scott C. Farrow, Khoa Chung, Richard M. E. Payne, Trinh-Don Nguyen, Thu-Thuy T. Dan, Inês Soares Teto Carqueijeiro, Konstantinos Koudounas, Thomas Dugé de Bernonville, Belinda Ameyaw, D. Marc Jones, Ivo Jose Curcino Vieira, V, S. E. O. (2018). Missing enzymes in the biosynthesis of the anticancer drug vinblastine in Madagascar periwinkle. *Science*, *360*(6394), 1235–1239. https://doi.org/10.1126/science.aat4100

M. Serif, B. Lepetit, K. Weißert, P.G. Kroth, C. R. B. (2017). A fast and reliable strategy to generate TALEN-mediated gene knockouts in the diatom Phaeodactylum tricornutum. *Algal Research*, *23*, 186–195. https://doi.org/10.1016/j.algal.2017.02.005

1101 Maury, J., Asadollahi, M. A., Møller, K., Schalk, M., Clark, A., Formenti, L. R., & Nielsen, J.
1102 (2008). Reconstruction of a bacterial isoprenoid biosynthetic pathway in Saccharomyces
1103 cerevisiae. *FEBS Letters*, *582*(29), 4032–4038. https://doi.org/10.1016/j.febslet.2008.10.045

1104 McBride, A. A. (2008). Chapter 4 Replication and Partitioning of Papillomavirus Genomes.
1105 *Advances in Virus Research*, *72*(08), 155–205. https://doi.org/10.1016/S0065-3527(08)00404-1

1106 Meyer, P. (1995). Understanding and controlling transgene expression. *Trends in Biotechnology*,
1107 *13*(9), 332–337. https://doi.org/10.1016/S0167-7799(00)88977-5

1108 Moritz, B., Woltering, L., Becker, P. B., & Göpfert, U. (2016). High levels of histone H3 acetylation
1109 at the CMV promoter are predictive of stable expression in Chinese hamster ovary cells.
1110 *Biotechnology Progress*, *32*(3), 776–786. https://doi.org/10.1002/btpr.2271

1111 Nicholls, P. K., Bellott, D. W., Cho, T. J., Pyntikova, T., & Page, D. C. (2019). Locating and
1112 characterizing a transgene integration site by nanopore sequencing. *G3: Genes, Genomes,*
1113 *Genetics*, *9*(5), 1481–1486. https://doi.org/10.1534/g3.119.300582

1114 Norrander, J., Kempe, T., & Messing, J. (1983). Construction of improved Ml3 vectors using
1115 oligodeoxynucleotidedirected mutagenesis (Ml3 cloning; synthetic DNA primers; gene
1116 synthesis; phosphoramidites; DNA polymerase I Klenow frag-ment). *Bren Road East*, *26*(612),
1117 935–7335.

1118 Nymark, M., Sharma, A. K., Sparstad, T., Bones, A. M., & Winge, P. (2016). A CRISPR/Cas9
1119 system adapted for gene editing in marine algae. *Scientific Reports*, *6*(April), 24951.
1120 https://doi.org/10.1038/srep24951

1121 Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., … Newman, J. D.
1122 (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*,
1123 *496*(7446), 528–532. https://doi.org/10.1038/nature12051

1124 Papapetrou, E. P., & Schambach, A. (2016). Gene insertion into genomic safe harbors for human
1125 gene therapy. *Molecular Therapy*, *24*(4), 678–684. https://doi.org/10.1038/mt.2016.38

1126 Parayil Kumaran Ajikumar, Wen-Hai Xiao, Keith E. J. Tyo, Yong Wang, Fritz Simeon, Effendi
1127 Leonard, Oliver Mucha, Too Heng Phon, Blaine Pfeifer, G. S. (2010). Isoprenoid Pathway
1128 Optimization for Taxol Precursor Overproduction in Escherichia coli. *Science*, *330*(6000), 70–
1129 74. https://doi.org/10.1038/nprot.2006.474

1130 Parker, J. D., Rabinovitch, P. S., & Burmer, G. C. (1991). Walking Polymerase Chain Reaction.
1131 *Gene*, *19*(11), 3055–3060.

1132 Pinder, M. I. M., Johansson, O. N., Almstedt, A., Kourtchenko, O., Clarke, A. K., Godhe, A., &
1133 Töpel, M. (2019). Genome Sequence of Kordia sp. Strain SMS9 Identified in a Non-Axenic
1134 Culture of the Diatom Skeletonema marinoi . *Journal of Genomics*, *7*, 46–49.
1135 https://doi.org/10.7150/jgen.35061

1136 Pinto, F., Pacheco, C. C., Oliveira, P., Montagud, A., Landels, A., Couto, N., … Tamagnini, P.
1137 (2015). Improving a Synechocystis-based photoautotrophic chassis through systematic genome
1138 mapping and validation of neutral sites. *DNA Research*, *22*(6), 425–437.

https://doi.org/10.1093/dnares/dsv024

Pollak, B., Matute, T., Nunez, I., Cerda, A., Lopez, C., Kan, A., … Roscoff, S. B. De. (2019). Universal Loop assembly (uLoop): open, efficient, and species- agnostic DNA fabrication.

Rajeevkumar, S., Anunanthini, P., & Sathishkumar, R. (2015). Epigenetic silencing in transgenic plants. *Frontiers in Plant Science*, *6*(September), 1–8. https://doi.org/10.3389/fpls.2015.00693

Remmers, I. M., D'Adamo, S., Martens, D. E., de Vos, R. C. H., Mumm, R., America, A. H. P., … Lamers, P. P. (2018). Orchestration of transcriptome, proteome and metabolome in the diatom Phaeodactylum tricornutum during nitrogen limitation. *Algal Research*, *35*(August), 33–49. https://doi.org/10.1016/j.algal.2018.08.012

Robertsen, E. M., Kahlke, T., Raknes, I. A., Pedersen, E., Semb, E. K., Ernstsen, M., … Willassen, N. P. (2016). META-pipe - Pipeline Annotation, Analysis and Visualization of Marine Metagenomic Sequence Data, (1), 1–22.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *OECD Observer*, *29*(220), 23–24. https://doi.org/10.1038/nbt0111-24

Salsman, J., & Dellaire, G. (2016). Precision Genome Editing in the CRISPR Era.

Sau, S., Ghosh, S. K., Liu, Y. T., Ma, C. H., & Jayaram, M. (2019). Hitchhiking on chromosomes: A persistence strategy shared by diverse selfish DNA elements. *Plasmid*, *102*(January), 19–28. https://doi.org/10.1016/j.plasmid.2019.01.004

Scaife, M. a., & Smith, a. G. (2016). Towards developing algal synthetic biology. *Biochemical Society Transactions*, *44*, 716–722. https://doi.org/10.1042/BST20160061

Serif, M., Lepetit, B., Weißert, K., Kroth, P.G., & Rio Bartulos, C. (2017). A fast and reliable strategy to generate TALEN-mediated gene knockouts in the diatom Phaeodactylum tricornutum. *Algal Researc*h, 23, 186–195. https://doi.org/10.1016/j.algal.2017.02.005

Serif, M., Dubois, G., Finoux, A. L., Teste, M. A., Jallet, D., & Daboussi, F. (2018). One-step generation of multiple gene knock-outs in the diatom Phaeodactylum tricornutum by DNA-free genome editing. *Nature Communications*, *9*(1), 1–10. https://doi.org/10.1038/s41467-018-06378-9

Sharma, A. K., Nymark, M., Sparstad, T., Bones, A. M., & Winge, P. (2018). Transgene-free genome editing in marine algae by bacterial conjugation – comparison with biolistic CRISPR/Cas9 transformation. *Scientific Reports*, *8*(1), 14401. https://doi.org/10.1038/s41598-018-32342-0

Sheff, M. A., & Thorn, K. S. (2004). Optimized cassettes for fluorescent protein tagging in Saccharomyces cerevisiae. *Yeast*, *21*(8), 661–670. https://doi.org/10.1002/yea.1130

Shin, S.-E., Lim, J.-M., Koh, H. G., Kim, E. K., Kang, N. K., Jeon, S., … Jeong, B. (2016). CRISPR/Cas9-induced knockout and knock-in mutations in Chlamydomonas reinhardtii SUPP. *Scientific Reports*, *6*(April), 27810. https://doi.org/10.1038/srep27810

1176 Slattery, S. S., Diamond, A., Wang, H., Therrien, J. A., Lant, J. T., Jazey, T., … Edgell, D. R. (2018).
1177 An Expanded Plasmid-Based Genetic Toolbox Enables Cas9 Genome Editing and Stable
1178 Maintenance of Synthetic Pathways in *Phaeodactylum tricornutum*. *ACS Synthetic Biology*,
1179 acssynbio.7b00191. https://doi.org/10.1021/acssynbio.7b00191

1180 Smetacek, V. (1999). Diatoms and the Ocean Carbon Cycle. *Protist*, *150*(1), 25–32.
1181 https://doi.org/10.1016/S1434-4610(99)70006-4

1182 Smith, S. R., Dupont, C. L., McCarthy, J. K., Broddrick, J. T., Oborník, M., Horák, A., … Allen, A.
1183 E. (2019). Evolution and regulation of nitrogen flux through compartmentalized metabolic
1184 networks in a marine diatom. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-
1185 019-12407-y

1186 Stukenberg, D., Zauner, S., Aquila, G. D., & Maier, U. G. (2018). Optimizing CRISPR / Cas9 for the
1187 Diatom Phaeodactylum tricornutum, *9*(June), 1–11. https://doi.org/10.3389/fpls.2018.00740

1188 Tanwar, A., Sharma, S., & Kumar, S. (2018). Targeted genome editing in algae using CRISPR/Cas9.
1189 *Indian Journal of Plant Physiology*, *23*(4), 1–17. https://doi.org/10.1007/s40502-018-0423-3

1190 Van Moerkercke, A., Fabris, M., Pollier, J., Baart, G. J. E., Rombauts, S., Hasnain, G., … Goossens,
1191 A. (2013). CathaCyc, a metabolic pathway database built from catharanthus roseus RNA-seq
1192 data. *Plant and Cell Physiology*, *54*(5), 673–685. https://doi.org/10.1093/pcp/pct039

1193 Vavitsas, K., Fabris, M., & Vickers, C. E. (2018). Terpenoid Metabolic Engineering in, (Figure 1).
1194 https://doi.org/10.3390/genes9110520

1195 Wang, C., Zada, B., Wei, G., & Kim, S. W. (2017). Metabolic engineering and synthetic biology
1196 approaches driving isoprenoid production in Escherichia coli. *Bioresource Technology*, *241*,
1197 430–438. https://doi.org/10.1016/j.biortech.2017.05.168

1198 Weyman, P. D., Beeri, K., Lefebvre, S. C., Rivera, J., Mccarthy, J. K., Heuberger, A. L., … Dupont,
1199 C. L. (2015). Inactivation of Phaeodactylum tricornutum urease gene using transcription
1200 activator-like effector nuclease-based targeted mutagenesis. *Plant Biotechnology Journal*, *13*(4),
1201 460–470. https://doi.org/10.1111/pbi.12254

1202 Yao, Y., Lu, Y., Peng, K. T., Huang, T., Niu, Y. F., Xie, W. H., … Li, H. Y. (2014). Glycerol and
1203 neutral lipid production in the oleaginous marine diatom Phaeodactylum tricornutum promoted
1204 by overexpression of glycerol-3-phosphate dehydrogenase. *Biotechnology for Biofuels*, *7*(1), 1–
1205 9. https://doi.org/10.1186/1754-6834-7-110

1206 Yasuyuki Nakamura, A., Teruyuki Nishi, a, B., Risa Noguchi, c Yoichiro Ito, a Toru Watanabe, B.,
1207 Tozo Nishiyama, b Shimpei Aikawa, a * Tomohisa Hasunuma, a Jun Ishii, a Yuji Okubo, B., &
1208 Akihiko Kondoa, D. (2018). A Stable, Autonomously Replicating Plasmid Vector Containing
1209 Pichia pastoris Centromeric DNA, 1–16.

1210 Yen-Ting-Liu1, Saumitra Sau1, Chien-Hui Ma1, Aashiq H Kachroo1, Paul A Rowley1, Keng- Ming
1211 Chang1, Hsiu-Fang Fan2, and M. J. (2008). The partitioning and copy number control systems
1212 of the selfish yeast plasmid: an optimized molecular design for stable persistence in host cells.
1213 *Bone*, *23*(1), 1–7. https://doi.org/10.1038/jid.2014.371

1214 Zurbriggen, A., Kirst, H., & Melis, A. (2012). Isoprene Production Via the Mevalonic Acid Pathway
1215 in Escherichia coli (Bacteria). *Bioenergy Research*, *5*(4), 814–828.
1216 https://doi.org/10.1007/s12155-012-9192-4

1217